Representation and Processing of Chinese Nominals and Compounds

Evelyne Viegas, Wanying Jin, Ron Dolan and Stephen Beale
New Mexico State University
Computing Research Laboratory
Las Cruces, NM 88003, USA
viegas,wanying,ron,sb@crl.nmsu.edu

Abstract

[1]In this paper, we address representation issues of Chinese nominals. In particular, we look at lexical rules as a conceptual tool to link forms with the same semantics as is the case between nominalisations and the forms they are derived from. We also address Chinese compounds, illustrating how to recover implicit semantic relations in nominal compounds. Finally, we show how to translate Chinese nominals within a knowledge-based framework.

1 Introduction

In this paper, we present results of a theoretical and an applied investigation, within a knowledge base framework, on the building and processing of computational semantic lexicons, as reflected by experiments done on Spanish, English and Chinese, with a large scale application on Spanish. The multilingual dictionaries making process (Viegas and Raskin, 1998) has been tested and attested for **Mikrokosmos**, a machine translation system (Nirenburg *et al.*, 1996) from Spanish and Chinese to English.[2] Here, we focus on Chinese nominals and compounds in terms of representation and processing.

In Section2, we briefly present the information carried inside Mikrokosmos lexicons. In Section 3, we show how a semantic-based transcategorial approach is best fitted to account for nominalisations and their derived forms. Formally, we use the conceptual tool of lexical rules as described in (Viegas *et al.*, 1996). In Section 4, we address the translation of Chinese nominal compounds into English using semantic information and word order information. We show the advantage of a transcategorial approach to lexicon representation and investigate some trade-offs between an interlingua and transfer approach to nominal compounding.

[1]This work has been supported in part by DoD under contract number MDA-904-92-C-5189.

[2]The interested reader can visit the Mikrokosmos site at http://crl.nmsu.edu/Research/Projects/mikro/.

2 A Brief Overview on the Structure of Mikrokosmos Lexicons

In Mikrokosmos, the lexical information is distributed among various levels, relevant to phonology, orthography, morphology, syntax, semantics, syntax-semantic linking, stylistics, paradigmatic and syntagmatic information, and also database type management information.[3]

Each entry consists of a list of words, stored in the lexicon independently of their POS (the verb and noun form of *walk* are under the same superentry). Each word meaning is identified by a unique identificator, or lexeme (Onyshkevych and Nirenburg, 1994). Homonyms and all meaning shifts of polysemous words are listed under one single superentry.[4]

We illustrate in Figure 1 relevant aspects, for this paper, of a lexicon entry via the description of two senses of the Chinese word 活动 (activity): WorkActivity and Exercise, which are well defined symbols or concepts in the Mikrokosmos ontology as described in (Mahesh, 1996).

Word meanings in Mikrokosmos are represented partly in the lexicon and partly in the ontology. We have strived to achieve an intermediate grain size of meaning representation in both the lexicon and the ontology: many word senses have direct mappings to concepts in the ontology; many others must be decomposed and mapped indirectly through composition and modification of ontological concepts. We have developed a set of guidelines and a training methodology that results in acceptable quality and uniformity in lexical and ontological representations (Mahesh, 1996; Viegas and Raskin, 1998). In principle, the separation between ontology and lexicon is as follows: language-neutral meanings are stored in the former; language-specific information in the latter.

We keep the number of concepts well below the number of lexical items for a given language, such

[3]Details on these zones can be found in (Viegas and Raskin, 1998; Meyer *et al.*, 1990).

[4]See (Weinreich, 1964), (Fillmore, 1971), (Cruse, 1993), (Pustejovsky, 1995) for interesting accounts on homography/polysemy. See (Bouillon *et al.*, 1992), (Busa, 1996) for accounts on nominals.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1998		**00-00-1998 to 00-00-1998**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Representation and Processing of Chinese Nominals and Compounds	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Department of Computer Science,New Mexico State University,PO Box 30001,Las Cruces,NM,88003-8001**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**5**	

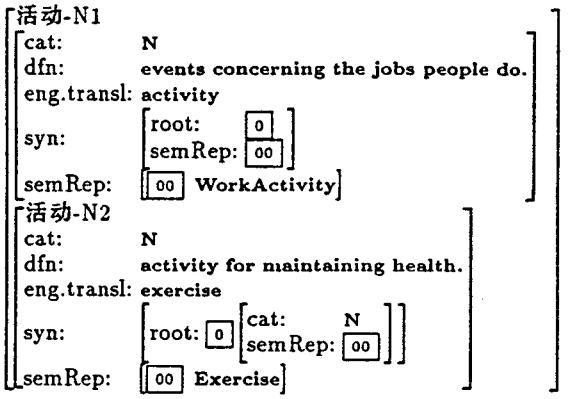

Figure 1: Partial Entry for the Chinese word 活动.

that, for instance, the concept Ingest can be lexicalised as 吃 (*eat*) or 喝 (*drink*) according to the constraints put in the lexicon on the theme: Food for *eat* and Liquid for *drink* respectively.

3 Role of Lexico-Semantic Rules

This section deals with the use of morpho-semantic lexical rules (MSLRs) in the process of large-scale acquisition. The advantage of MSLRs is twofold: first, they can be considered as a means to reduce the number of lexicon entry types, and generally to make the acquisition process faster and cheaper; second, they can enhance the results of analysis processing by creating new entries for unknown words from the lexicon, found in corpora. Lexical rules have been addressed by many researchers. Here we apply Viegas *et al.* (1996) methodology to Chinese.

Briefly, applying MSLRs to the Spanish entry *comprar* (buy), our MSLR generator produced automatically 26 new entries (*comprador-N1* (buyer), *comprable-Adj* (buyable), etc). This includes creating new syntax, semantics and syntax-semantic mappings with correct subcategorisations and also the right semantics; for instance, the lexical entry for *comprable* will have the subcategorisation for predicative and attributive adjectives and the semantics adds the attribute "FeasibilityAttribute" to the basic meaning "Buy" of *comprar*. The form list generated by the morpho-semantic generator is checked against MRDs, dictionaries and corpora: only the forms found in them are submitted to the acquisition process.

MSLRs constitute a powerful conceptual tool to extend a core lexicon from a monolingual viewpoint. We applied the same methodology to Chinese. The rules are language independent, what is language dependent is the morphemes to which they can apply. For Chinese, we do not have to worry about developing a morpho-semantic generator as the productivity in morphology is poor, if one excepts compounds characters in which semantics is not compositional (see the example of 葡萄糖 (glucose) below). In this latter case, we acquired the entry manually. So rules are used to link nominalisations to verbs, and vice versa, meaning that once verbs have been acquired, nominal derivations can be produced automatically using rules.

(1) a. 肯定-V1(affirm)
    b. 肯定-N1(affirmation)

We present below the entry for 肯定-N1 (affirmation) after the application of the LR2event rule on 肯定-V1 (affirm).

```
#0= [key:肯定, gram:[pos: N], semRep:#sem,
    synSem:[gram:#o1, semRep:#t], [gram:#o2, semRep:#a],
    lexRule: LR2event[root:[key:肯定,
                gram:[pos:V, subc:NPVNP],
                semRep:#sem=[name:Assert, agent:#a,
                        theme:#t]], vn:#0]]
```

In our corpus (from the Chinese newspaper Xinhua Daily), we found that 166 nouns could be derived from 351 verbs; that is, almost 47% of verbs can produce nouns. From an acquisition viewpoint, it is cheaper to use the mechanism of lexical rules to automatically produce nouns from verbs, with the same semantics, this is due to our transcategorial approach to semantics, where the same piece of semantics can be lexicalised as either a Noun or Verb.

4 A Transcategorial Approach

Compounding in Chinese is a common phenomenon (Jin, 1994; Jin and Chen, 1995; Palmer and Wu, 1995). It is mainly used to combine i) characters whose semantics is different and non compositional, and ii) sequences of nouns.

In i) we create entries for single characters and entries for combined characters, (e.g., (2)):

(2) a. 葡萄 (grape)
    b. 糖 (sugar)
    c. 葡萄糖 (glucose)

In this paper we are concerned with ii) only. In the following, we investigate three ways to translate Chinese nominal compounds into English, using word order information, semantic information and co-occurring information in syntactic, semantic and transfer approaches, respectively.

4.1 A Syntactic Approach

Compounds proliferate in Chinese. The head of the compound seems to be easily identified as the last noun in the sequence, and therefore in the task of translating Chinese compounds into English compounds, where English also makes use of compounds as opposed to say French, one could adopt a transfer-based approach, where each Chinese noun is translated into English in the same sequence: 应用 软件 (*application software*); 资料 管理 系统 (*data management system*). It gets a bit more complex

when there is a large sequence of nouns in English, whereas it is still acceptable and normal in Chinese. In our corpus we found compounds containing up to 6 nouns: 军事 理论 考核 题库 管理 系统 (military theory test database management system) (*the management system of database for testing military theory*). In these cases, it is difficult to comprehend the compound in English and some "linking information" is needed to break the compound and make it understandable in English. This is where the semantics comes in, as one needs to understand the underlying relationships between the nouns, and identify "sub-heads" inside the Chinese compounds, which will become the heads of English smaller compounds linked via relations. For instance, in 军事 理论 考核 题库 管理 系统 (military theory test database management system)(*the management system of database for testing military theory*) one might want to "break" the Chinese compound into smaller English compounds "management system," " database" and "military theory" with a relation "test" between the last compounds (*the management system of database for testing military theories*).

4.2 A Lexico-Semantic Approach

We now show examples of how the semantics can help identify sub-heads inside the Chinese compound (the head of the Chinese compound is the last noun). Second, we show how a transcategorial approach can help go from an NN compound 经济 政策 (economy policy) in Chinese to AdjN constructions (*economic policy*) in English. Finally, we show how nominal mismatches are dealt with as a generation issue. For illustration purposes, we will mainly consider compounds composed of two nouns; however, this semantic approach applies to more than 2 nouns. Lexemes can be mapped to Objects (O) ("Car" *car*), Events (E) ("Explode" *explosion*), Relations (R) ("Utilizes" *use*) or Attributes (A) ("ColourAttribute" *colour*). In the case of NNs, we have 14 combinations allowed (RR and AA do not seem to co-occur), where E, O and R can be heads, with the following hierarchy of headhood:

$$E > R > O$$

When the semantics of the NN is expressed with a combination of identical types (e.g. EE or OO), the semantic analyser must score the constraints between the two nouns to find the head. Sometimes it is possible to find the semantic relation linking the two nouns in the ontological entry of the nouns, as in the example OO below.

(OO) Object - Object

```
[np [mod 计算机 (n, ji4suan4ji1, computer, Computer)]
    [n   技术   (n, ji4shu4, technology, Technology)]]
```

Here, both nouns are typed as O, and therefore we need a mechanism to assign the head. The generator must identify the underlying relation between the Os. This can be done by searching for a relation R in the ontology shared by the 2 Os, such as "applied-to" with a domain which is in an ISA relationship with "technology" and a range also in an ISA relationship with "computer". Needless to say that this approach is knowledge intensive, and in case we do not have this type of knowledge encoded we rely on a transfer-based approach following the Chinese word order. Here, we could successfully generate *technology about computer* and *computer technology*, with a preference on the latter.

(OR) Object - Relation

```
[np [mod 行 (n, hang2, business, AreaOfExpertise)]
    [n   长 (n, zhang3, leader, HeadOf)]]
```

"HeadOf" is a relation and therefore the head, as the other noun is an O. The generator can lexicalise this as *leader of business* or *business leader* via a rule; the latter is assigned a preference in absence of modifiers such that we can still generate *the leader of a big business* instead of *big business leader*. Note that we do not need to use the hierarchy in the case of only two Ns to identify the head because, the head is the last noun in a Chinese compound; we showed this example, in case it entered in a larger compound such as "business leader major office" where one might want to break it as "the major office" of "the business leader".

(EA) Event - Attribute

```
[np [mod 工作 (n, gong1zuo4, work, WorkActivity)]
    [n   方式 (n, fang1shi4, style, StyleAttribute)]]
```

Here, E is the head and this semantics is lexicalised as *way of working* or *work style*, with again a preference on the latter.

(OR) illustrates our transcategorial approach:

```
[np [mod 经济 (n, jing1ji4, economy, Economy)]
    [n   效益 (n, xiao4yi4, benefit, BenefitFrom)]]
```

Here is a case where our transcategorial approach to lexicon representation helps in generating an AdjN construction *economic benefit* for an NN Chinese compound; this is due to the fact that both *economy* and *economic* share the same semantics, and thus the generator will present both possibilities; moreover, they co-occur in English whereas *economy* and *benefit* do not. The head is easily identified in R "BenefitFrom" and as such the compound could also be generated as *benefit to economy*.

(OEE) illustrates a phrase

```
[np [mod
    [mod 科技 (abbr, ke1ji4, science&technology, Science)]
    [n   攻关 (n, gong1guan1, attack-key-problem, Solve+Att)]]
    [n 计划 (n, ji4hua4, plan, PlanningEvent)]]
```

This NNN compound presents a case of mismatch between Chinese and English, it can be paraphrased as: *plan to solve key problems in science and technology*. Here, a transfer-based approach would fail to translate adequately, as 攻关 (attack-key-problem) must be expressed as an expression equivalent to *solving important problems*, and as such the following English compound *science technology solving key problem plans* must be broken into smaller compounds with explicit relations between them.[5]

These examples illustrate why a semantic approach is preferable, and sometimes necessary, to translate Chinese compounds into English. However, as discussed earlier, 1) this approach is knowledge intensive, and 2) English compounding seems to follow the same Chinese word order regularly enough so that we consider using a transfer approach as a back-up to generation.

4.3 A Transfer-based Approach

Semantics can be expensive to use so we also rely on a transfer-based approach as a back-up method when semantics fails to give us the semantic relation between the nouns. We can do this because English allows compounding (whereas for French and Spanish, a transfer approach would be more problematic as compounding is not as productive and relations must be identified). However, as we noticed previously, it can become difficult in English to get the meaning of a large compound, it is therefore better to "break" the compound into 2 or 3 compounds. We hope to bypass part of this problem by using co-occurring information in a transfer approach.[6]

```
[np
 [mod 计算机 (n, ji4suan4ji1, computer, Computer)]
 [mod
  [mod
   [mod 军事 (n, jun1shi4, military-affairs, MilitaryActivity)]
   [n 理论 (n, li3lun4, theory, Theory)]]
  [n 考核 (n, kao3he2, test, Examination)]]
 [n 题库 (n, ti2ku4, database, Database)]]
```

computer database for test of theory of military affairs

In this case, only co-occurring information will signal the generator to link "computer" to "database" to produce "computer database"; this information must be encoded in the lexicon, as we show in next section.

```
[np
 [mod
  [mod
   [mod
    [mod 军事 (n, jun1shi4, military-affairs, MilitaryActivity)]
    [n 理论 (n, li3lun4, theory, Theory)]]
   [n 考核 (n, kao3he2, test, Examination)]] .
  [n 题库 (abbr, ti2ku4, text database, Database)]]
 [np
  [mod 管理 (n, guan3li3, management, ManagementActivity)]
  [n 系统 (n, xi4tong3, system, System)]]]
```

Following the Chinese word order seems to be acceptable in English, to produce *military theory test database management system*. However, a better translation might be *the management system of a database for testing military theory*, in which case, relations between nouns must be made explicit, using the semantic information found in the ontological concepts in a semantic approach.

5 Processing of Chinese Nominals and Nominal Compounds

We utilise an efficient constraint-based control mechanism called *Hunter-Gatherer* (HG) (Beale, 1997) to process Chinese nominals and compounds. This mechanism has been successfully applied to the analysis of Spanish and generation of English. We refer to (Beale *et al.*, 1995) for details on how the semantic analyser works, and (Beale *et al.*, 1997) on how the generator works.

In this paper, we are interested in showing how HG allows us to mark certain compositions as being dependent on each other: once we have two lexicon entries that we know go together, from either syntactic, lexical, or semantic viewpoints, HG will ensure that they are correctly treated. HG gives preference to words which "co-occur" together, from any of the above viewpoints. The analyser simply needs to detect the co-occurrence and add the constraint that the corresponding senses be used together.

In the case of "computer database," the lexicon entry for "database" encodes the syntagmatic relation (LSFSyn) which keeps the semantics of the nouns compositional and signals the processor (analyser or generator) to consider the nouns as syntactically linked:

```
#0=[key: "题库",
   rel: [syntagmatic: LSFSyn [base: #0,
        co-occur: [key: "计算机", sense: n1, ...]]]]
```

We provide below the example of a Chinese sentence, its English translation and relevant parts of the result of the semantic analysis, showing the analysis of the compound 攻关 "tackle-key-problem".

Chinese Sentence Example 这个攻关大项由国家海洋局主持全国34个单位的20000多名科技人员参加攻关，是一个包括7个课题，39个专题的大型工程性应用项目。

[5] In "Solve+Attitude", Attitude reflects the importance attached to the event.

[6] We saw that in a semantic approach the headhood hierarchy provides a good clue to break a compound.

English Literal Translation This classifier attack-key-problem big project by State-Maritime-Bureau direct whole country 34 classifier units adjective-marker 2000 more classifier scientific-and-technical personnel participate attack-key-problem, is one classifier including 7 classifier tasks, 39 classifier special topics adjective-marker large-scale engineering application project.

English Translation This project which deals with important problems, directed by the State Maritime Bureau and in which participated more than 2,000 scientific and technical personnel from 34 units throughout the country, was a large-scale engineering application project including seven tasks and 39 special topics.

Partial Text Meaning Representation

```
SOLVE-219
 LOCATION      : SPEAKER
 TIME          : SPEAKER-TIME
 RELATION      : RESEARCH-216
 STRENGTH-ATTRIBUTE: 0.9
 THEME         : PROBLEM-220
PROBLEM-220
 THEME-OF      : SOLVE-219
ATTITUDE-225
 ATTRIBUTED-TO  : SPEAKER
 ATTITUDE-VALUE : 1
 TYPE           : SALIENCY
 ATTITUDE-SCOPE : PROBLEM-220
```

6 Conclusions - Perspectives

In this paper, we showed the advantage of adopting a transcategorial (semantic-based) approach to relate verbs with their nominalisations. We showed how to use lexico-semantic rules to relate different forms carrying the same semantics. These rules can be applied at run time in analysis, thus facilitating a syntactico-semantic recovery for unknown words. Concerning compounds, we have shown that we cannot avoid a semantic approach if we want a high quality translation, because of the number of nouns which can enter into a Chinese compound making it difficult to get the meaning of the compound in English. Thus, breaking the compound necessitates an understanding of the Chinese compound. However, we have suggested transfer-like approach for Chinese to English translation with the use of co-occurrences to "signal" privileged lexical links (*computer database*).

We have illustrated that by considering the information in the lexicon as constraints, the linguistic difference between pure compositionality and co-occurrent information becomes a virtual difference for HG.

References

S. Beale, S. Nirenburg and K. Mahesh. 1995. Semantic Analysis in the Mikrokosmos Machine Translation Project. In Proc. of *the 2nd Symposium on NLP*, Bangkok.

S. Beale. 1997. *HUNTER-GATHERER: Applying Constraint Satisfaction, Branch-and-Bound and Solution Synthesis to Computational Semantics.* Ph.D. Diss., Carnegie Mellon University.

S. Beale, E. Viegas and S. Nirenburg. 1997. Breaking Down Barriers: The Mikrokosmos Generator. In Proc. of *the NLP Pacific Rim Symposium*, Phuket.

P. Bouillon, K. Bösefelt and G. Russell. 1992. Compounds Nouns in a Unification-Based MT System. In Proc. of *the 3rd ANLP*, Trento.

F. Busa. 1996. *Compositionality and the Semantics of Nominals.* PhD. Disser. Brandeis University.

A. Cruse. 1993. Towards a Theory of Polysemy. *Building Lexicons for Machine Translation.* TR. SS-93-2, Stanford: AAAI Press.

C. Fillmore. 1971. Types of Lexical Information. In *Steinberg and Jakobovitz.*

W. Jin and L. Chen. 1995. Identify Unknown Words in Chinese Corpus. In Proc. of *the 3rd NLP Pacific-Rim Symposium*, Vol. 1, Seoul.

W. Jin. 1994. Chinese Segmentation Disambiguation. In Proc. of *COLING*, Japan.

K. Mahesh. 1996. *Ontology Development: Ideology and Methodology.* TR. MCCS-96-292, NMSU, CRL.

I. Meyer, B. Onyshkevych and L. Carlson. 1990. *Lexicographic Principles and Design for KBMT.* TR. CMU-CMT-90-118, CMU.

S. Nirenburg, S. Beale, S. Helmreich, K. Mahesh, E. Viegas, and R. Zajac. 1996. Two principles and six techniques for rapid MT development. In Proc. of *the 2nd AMTA.*

M. Palmer and Z. Wu. 1995. Verb Semantics for English-Chinese Translation. In *Machine Translation*, Volume 10, Nos 1-2.

J. Pustejovsky. 1995. *The Generative Lexicon.* MIT Press.

E. Viegas, B. Onyshkevych, V. Raskin and S. Nirenburg. 1996. From *Submit* to *Submitted* via *Submission*: on Lexical Rules in Large-scale Lexicon Acquisition. In Proc. of *the 34th ACL*, CA.

E. Viegas and V. Raskin. 1998. *Computational Semantic Lexicon Acquisition: Methodology and Guidelines.* TR. MCCS-98-315. NMSU: CRL.

U. Weinreich. 1964. Webster's Third: A Critique of its Semantics. In *International Journal of American Linguistics* 30: 405-409.